The Search of Tradition and Modernity: Digital Humanity Method in Studying Natsume Soseki and His Novels

Zixuan (Armstrong) Li

History 582A: Digital Humanities Methods for History and Literature
Mark Ravina
December 8, 2016

Introduction

Natsume Soseki, born in 1867, the year before the Meiji Restoration, was a Japanese author and literary scholar. Soseki was an essayist, an expert in Japanese poetry, *haiku*, and Chinese poetry, *kanshi*, a translator of classical Japanese to English, a scholar in English literature and most famously a novelist. In his short but prolific career from 1905 to 1916, Soseki wrote fifteen novels, including one unfinished. Most of Soseki's novels depict the discomfort and confusion of people in the rapidly westernizing Japan. Soseki, like the typical characters in his stories, was perplexed between the idea of tradition and modernity. As a modern intellectual, Soseki was a proponent of westernization, but he could not avoid the solitude and agony in the changing Japanese society.

The westernization of Japan after the Meiji Restoration guaranteed material and military successes. The victories in the Sino-Japanese War, 1894-1895, and the Russo-Japanese War, 1904-1905, were indications that Japan had become a first-rate military power. Japanese enriched their country and strengthened their military, as the slogan of the Restoration said, *Fukoku Kyohei*, but was less successful in achieving what the other slogan suggested, *Bunmei Kaika*, civilization and enlightenment.[1] The inherently vague idea of *Bunmei Kaika* made it difficult to finish. The question was what kind of civilization and enlightenment should Japanese pursue. Many young Japanese took the idea of *Bunmei Kaika* simply as imitation of the western civilizations, but Soseki pointed out the superficial successes and understanding in the Japanese imitation. Soseki's study in Japan and abroad developed his keen perception on the idea of modernity and tradition. He

1 Isamu Fukuchi, "*Kokoro* and 'the Spirit of Meiji," *Monumenta Nipponica* 48.4, (1993): 469.

loved Chinese literature, but studied English in high school and college. He spent two unpleasant years in England from 1900 to 1901, indulging in studying and avoiding socializing, but kept writing about westerners in his travel to Manchuria and Korea.[2] Many of the characters in his works similarly suffer from the lost traditions and solitude, but Soseki concluded that these were inevitable price for being modern. Soseki loved virtues of both tradition and modernity, but he could not find a balance between them. His rejection of professorship in English literature, the ostensible forefront of modern study, caused sensation in Tokyo elites in 1907. Soseki justified his decision by confirming his disdain for the superficiality of Japanese academia. He worked afterwards for *Asahi* newspaper, where most of his novels were published.[3] Because of Soseki's importance in the early 20th century Japanese literature, western scholars produced numerous works about him. Through analyzing his works, people tried to understand Meiji Japan from its intellectuals and to build connections between Soseki's pursuit for modernity and Japanese endeavor to became a powerful state.

This paper attempts to use digital humanity methods to analyze western academia on Soseki, connecting them to textual data of his fifteen novels, and concludes with a close reading of his most famous work, *Kokoro (The Heart)*. The first part uses Data for Research (DfR) from JSTOR to study the English academia on Natsume Soseki. DfR has restrained ability to provide insight in

2 Sōseki Natsume, Inger Brodey and Sammy Tsunematsu, *Rediscovering, Natsume Sōseki: with the first English translation of Travels in Manchuria and Korea; celebrating the centenary of Soseki's arrival in England, 1900 - 1902* (Folkestone: Global Oriental, 2000): 1-10.

3 Edwin McClellan, *Two Japanese novelists: Sōseki and Tōson* (Chicago: University of Chicago Press, 1969): 3-15.

Japanese literature, since JSTOR holds limited amount of journals and most of them are in English. Nevertheless, the DfR search yields interesting results about the historiography of Japanese literature study. The second part analyzes Soseki's fifteen novels with the statistical computing language R and the Japanese tokenizer MeCab with the IPA dictionary. Major statistics are word frequency in percentage and term frequency-inverse document frequency (tf-idf), which measures distinctiveness of terms in a document. The third part combines quantitative analysis with a close reading of *Kokoro,* finding out what insights of statistics agree with perusing. The novel does not only possess resplendent aesthetic and symbolic values, but also has features that separate it from the rest of his works.

As many digital humanists point out, quantitative analysis in literature and history is not meant to provide clear answers and establish definitive theories. Economists and political scientists have been using statistical methods with caution for decades, because data may be misleading and may have different interpretations. An admonition that statisticians often give to student is that, "correlation does not imply causation". There are often some hidden factors in data, so making hasty conclusions is dangerous. In recent articles of digital humanists Dan Edelstein and Cameron Blevins, both authors asserted that studies in literature and history remain qualitative in nature. Distant reading cannot replace close reading, and people cannot frivolously use textual data as a proof.[4] This paper intends to observe what were the changing trends and focus in study of Natsume

4 Cameron Blevins, "Space, Nation, and the Triumph of Region: A View of the World from Houston," *Journal of American History* 101.1, (2014):126; Dan Edelstein, "Enlightenment Scholarship by the Numbers: dfr.jstor.org, Dirty Quantification, and the Future of the Lit Review," *Republic of Letters* Vol 4, Issue 1, (2014): 1-3.

Soseki. Historians and literary critics put much attention to Soseki's nature as an intellectual, who meditated on the idea of tradition and modernity. Text mining his novels points out that unrealistic love and inevitable solitude in the modernizing Japan are central themes in his works. A close reading of *Kokoro* illustrates how Soseki symbolized the modern trauma he experienced in Meiji Japan in this simple but beautiful work.

Historiographical Research with DfR from JSTOR

The historiographical research on Soseki utilizes DfR service from JSTOR. The search for the keyword "Natsume Soseki" returns a collection of 697 articles. DfR is not a perfect tool of conducting historiographical research, since it gives equal weight to an article about Shakespeare that mentions Soseki only once and a book review of Soseki's work. Two of the articles in the collection of 697 published in 2000 and 2001 mention the word "Shakespeare" around 3000 times. They are Shakespeare studies that merely mention Soseki. Nevertheless, these outliers do not seriously affect the result, since the collection is large. Figure 1, the key terms graph generated by DfR, shows that most frequent words in the collection is "Japanese" and "Japan". Other relevant

Figure 1. Key terms generated by DfR in the search for "Natsume Soseki"

words like “Meiji”, “modern”, “literature”, “translation”, “English” and “Chinese” are crucial terms in studies of Soseki’s works. Therefore, this collection of 697 documents are satisfactory to study the historiography about Soseki without causing severe bias. JSTOR is a database of predominantly English journals. In the 697 documents, 651 are in English (93%), 20 in Spanish, 13 in French, 8 in Italian and 5 in German. The little interruption of other languages is idea, since the paper only studies English terms.

There is no “Natsume” or “Soseki” in Figure 1. This could be problematic, if most of the 697 articles are about Japanese literature in general. Fortunately, as Figure 2 shows, terms “Natsume” and “Soseki” are more frequent than other Japanese authors. In the search for key terms in the DfR collection, the paper uses rolling mean of key terms frequency over five years to make smooth line graphs so that the changing trend is readable. Figure 2 presents the frequencies of the keywords “Natsume” and “Soseki” in thick black and blue lines, “Akutagawa”, “Dazai”, “Kawabata” and “Tanizaki”, four Japanese authors active in the 20th century, in warm color lines and “Chikamatsu”,

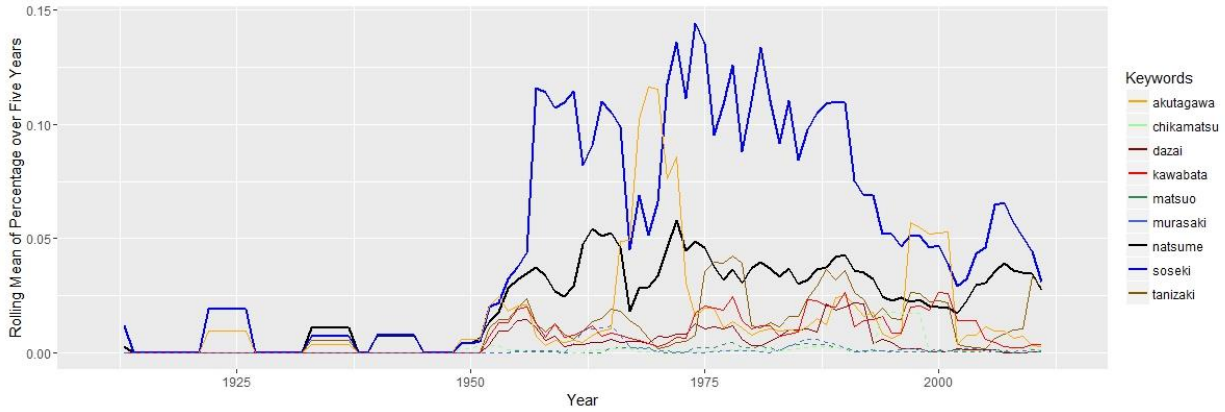

Figure 2. Rolling mean of frequencies of author names over five years from the collection

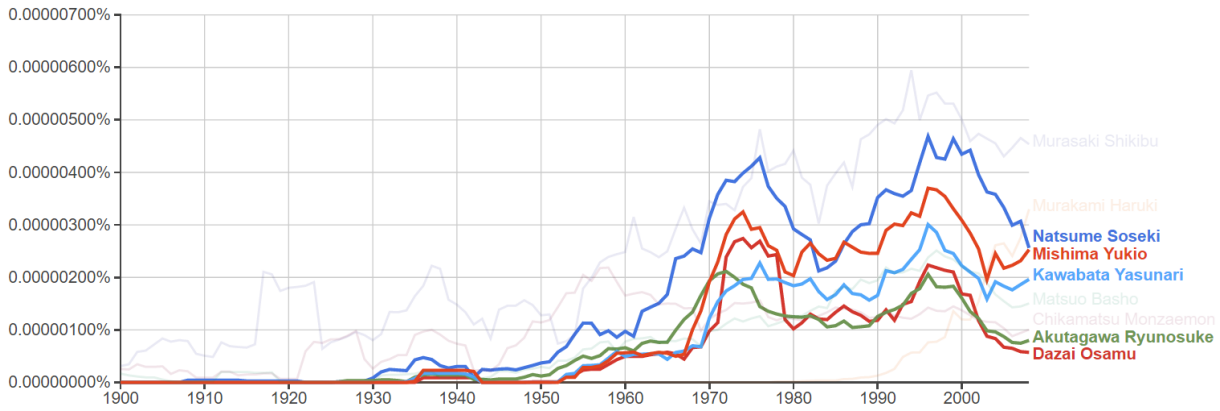

Figure 3. Frequencies of names of the early and mid-20th century Japanese authors in English books generated by Google Ngram

"Matsuo" and "Murasaki", three ancient Japanese authors, in cold color dashed lines. All the ancient authors stayed close to zero, while the modern authors appeared more. Two peaks of Akutagawa around1970 and 2000 means that Akutagawa and Soseki are strongly related in studies of English scholars. Akutagawa was, in fact, a disciple of Soseki. It is acceptable of the collection to mention other Japanese authors, especially who were active in the early and mid-20th century, since in Figure 3, Google Books displays that these writers appeared in the same trends in English books. Their frequencies followed the same patterns, with peaks in the 1970s and 1990s. The

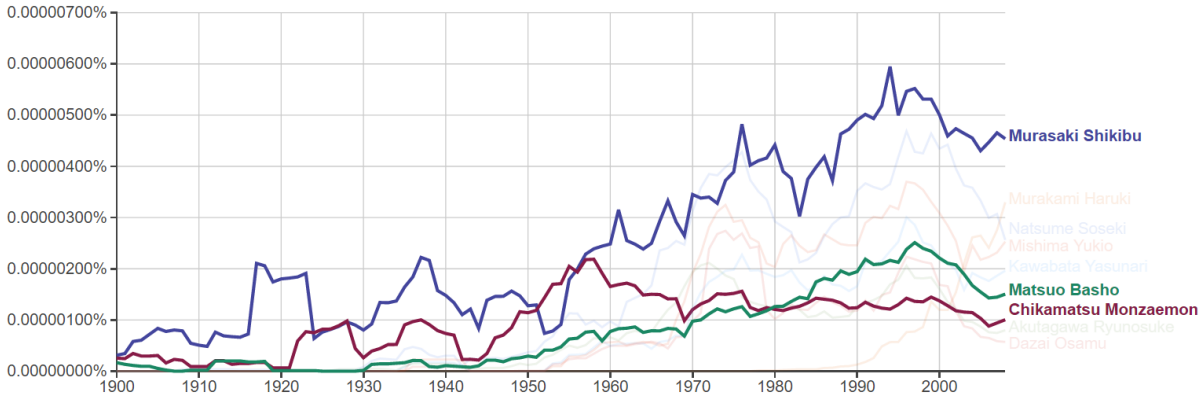

Figure 4. Frequencies of names of classical Japanese authors in English books generated by Google Ngram

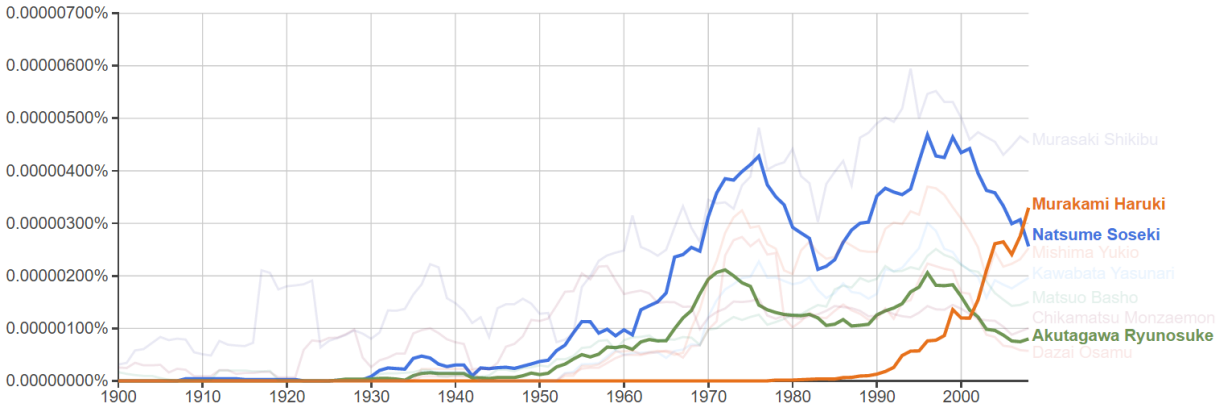

Figure 5. Frequency of Murakami Haruki in English books in comparison with Natsume Soseki and Akutagawa Ryunosuke generated by Google Ngram

searches for classical and contemporary Japanese authors did not follow this pattern. Figure 4 shows that Murasaki enjoyed a steadily rise of attention. Chikamatsu peaked in the 1950s, while Matsuo peaked in 1990s. Murakami's rise in Figure 5 is not surprising, since he attracted growing attention after his first work in 1982.

The nature of the collection of 697 article is clear. Most of them are studies of Japanese literature centered on Soseki, but also discuss other authors active in the early and mid-20th century.

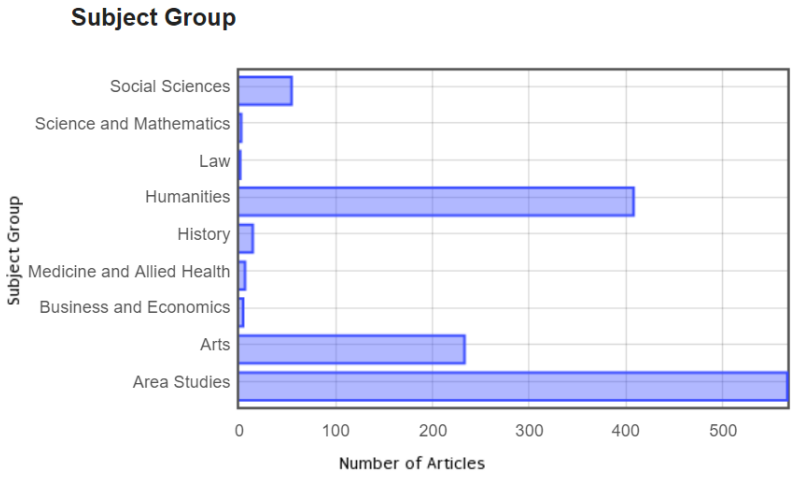

Figure 6. Subject groups of the collection generated by DfR from the collection

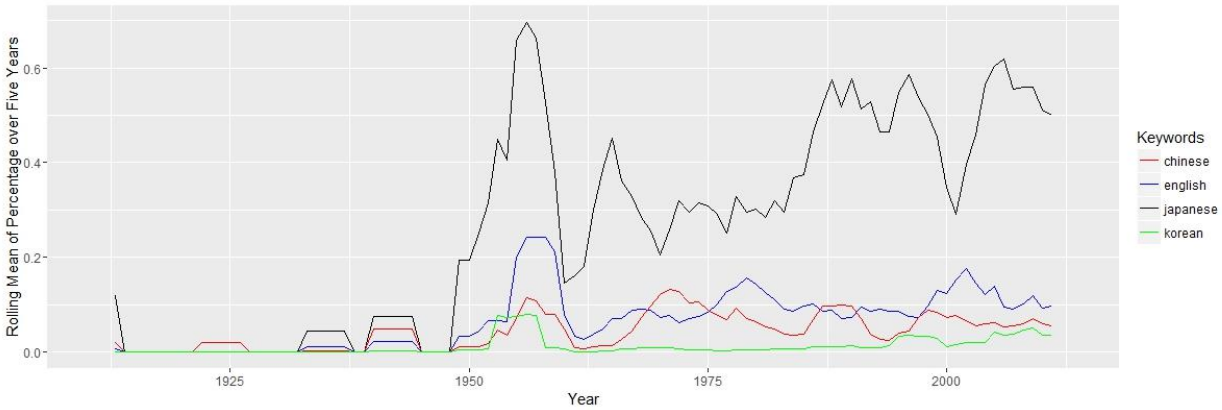

Figure 7. Rolling mean of frequencies of languages over five years from the collection

According to the subject groups in Figure 6, most articles fall in categories of area studies, arts and humanity. Articles in area studies are mostly in East Asian studies, as Figure 7 shows that the keyword “Chinese” is important as “English”. The word “Chinese” and “Korean” would not appear a lot, if English scholars were only interested in discussing Soseki’s works and their translations. Therefore, Soseki, a Japanese writer, was a lens for western academia to study East Asian cultures in general. Although Soseki was a modern writer and he appreciated liberty and individualism, Soseki remained his connections to the traditional East Asia. He loved Chinese poetry and painting; his novels often quoted classical Chinese; his pen name Soseki came from a

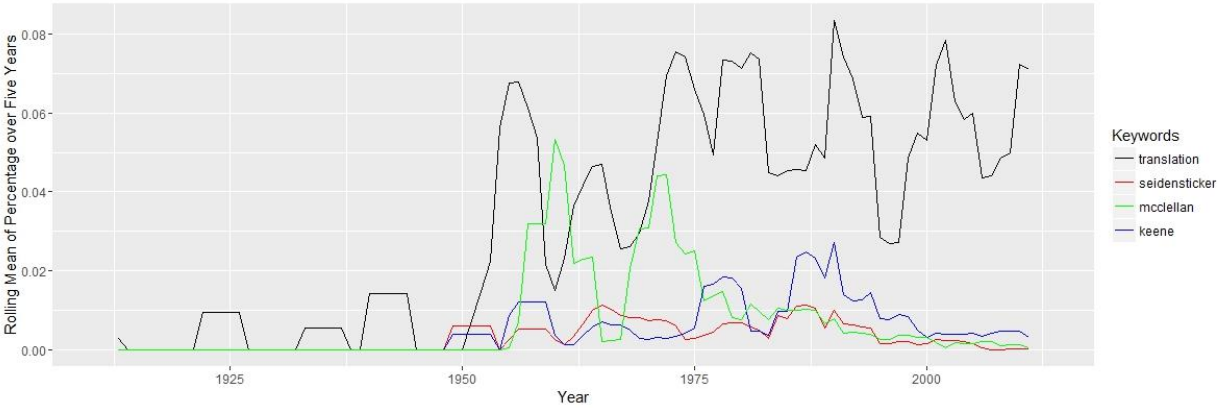

Figure 8. Rolling mean of frequencies of translation keywords over five years from the collection

Chinese tale *shushizhenliu (rinse with stone and sleep on stream)* in *Jinshu (The Book of Jin)*. The literary critic Eto Jun even commented that "Soseki stood with traditional morality, when he wrote the trauma of modern intellectuals in *Kokoro*".[5]

The continued attention of Soseki lead scholars to investigate of the connections in his works to East Asian cultural in general, but the early rise of Soseki's importance in western academia was a result of study of his works in translation. Figure 8 illustrates a boom of articles about Soseki after 1950. The black line shows the frequency of keyword "translation", and other lines are names of translators. Edwin McClellan, the British literary scholar and translator, introduced Soseki to the western audience and translated *Kokoro* and *Michikusa (Grass on the Wayside)*. The two peaks of the green line in Figure 8 around 1960 and 1970 corresponds to the rising frequency of the keyword "Soseki" in Figure 2 and "Japanese" in Figure 7. Western studies of Soseki from the 1950s to the 1970s centered around McClellan's translation. After 1975, McClellan's importance in studies of Soseki started to fall, since he researched on Mori Ogai and translated works of Shiga Naoya and Yoshikawa Eji. From the analysis of the frequency of the keywords, there was a clear transition from literary review of Soseki's works and their translations to a general East Asian study.

The next step is to examine how western scholars interpreted Soseki's works. In addition to their artistic values, Soseki's novels tell stories of Meiji Japanese, who were the interests of historians. Although Figure 6 shows that there were few articles strictly belong to history,

5 Fukuchi, "*Kokoro*", 480.

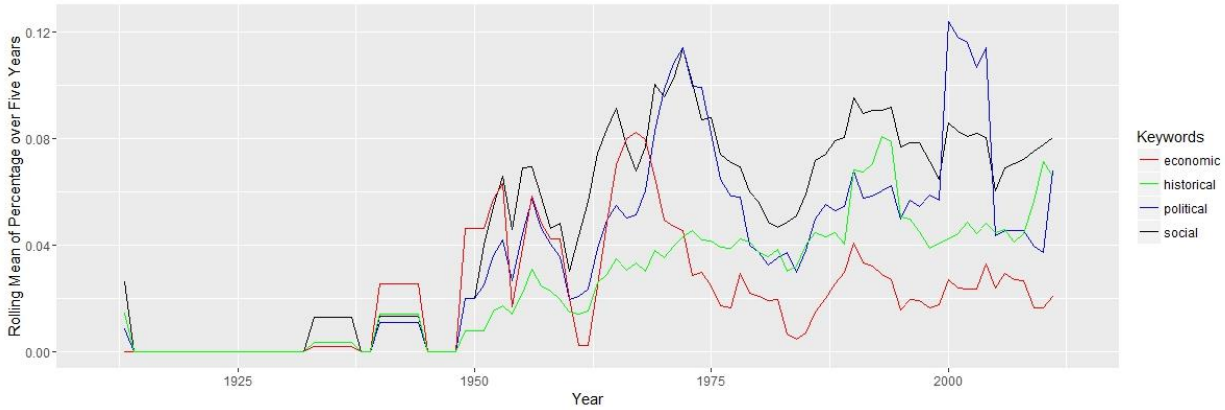

Figure 9. Rolling mean of frequencies of disciplinary keywords over five years from the collection

documents in area studies cannot discuss Soseki's work without noting its social, political, and historical importance. Figure 9 manifests the change of these four disciplinary keywords in frequency overtime. Meiji Japan was a period of expeditious social and political changes, which western scholars intended to find in Soseki's novels. These issues related to each other, since the line of "social" and "political" moved together. The line of "economic" moved with the "social" and "political" lines from 1950 to 1970, but its importance decreased since 1970. Economic conflicts and hardships are major plots in some of Soseki's novels. In *Kokoro*, the uncle defrauds Sensei of his inheritance, which leads to Sensei's isolation and distrust of relatives. Sensei exhorts

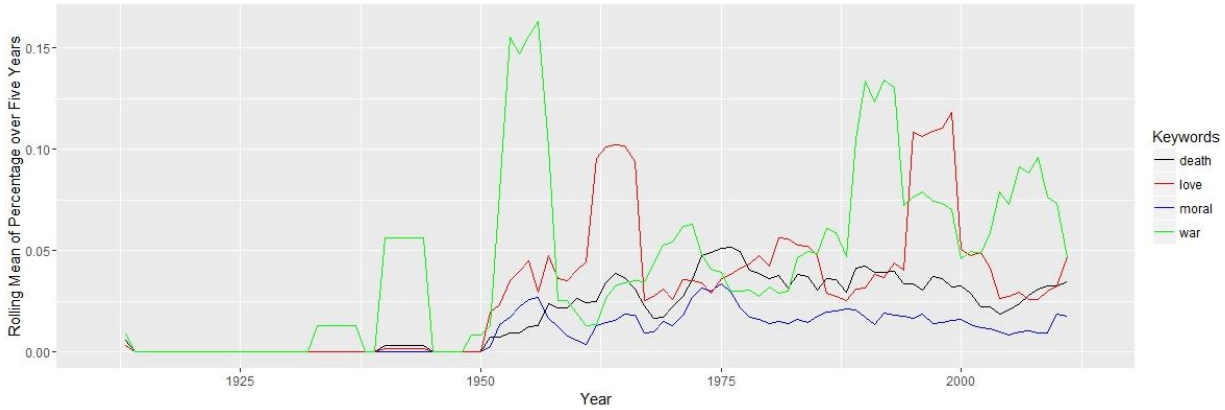

Figure 10. Rolling mean of frequencies of thematic keywords over five years from the collection

the young man not to believe his relatives, especially in financial problems. Western scholars, however, seemed to lose interest in economic implications in Soseki's novels. Financial issues might be essential concerns for individuals in Meiji Japan, but not for scholars interested in Japanese modernization. In contrast to the "economic" line, the "historical" line gained increasing attention. Before 1975 the word "historical" had less frequency than other three words, but after that its importance grew to the same level as the word "social" and "political". This might be a result of academic progress of Japanese history after the World War II. As Japanese westernization and wars faded away from contemporary politics to history, English scholars also became interested in placing Soseki's works in the historical context.

The high frequency of "war" in Figure 10 is a partial explanation to the interest of historical implications in Soseki's works. Although other keywords in the graph, "love", "death" and "moral" better described Soseki's works, the word "war" appeared to be more important in the discussions. The "war" line had two peaks around World War II, when other themes had frequency close two zero. In the 1950s, 1990s and 2000s, the frequency peaked again. The connection that western scholars built between Soseki and war was self-explanatory, since people were interested in finding

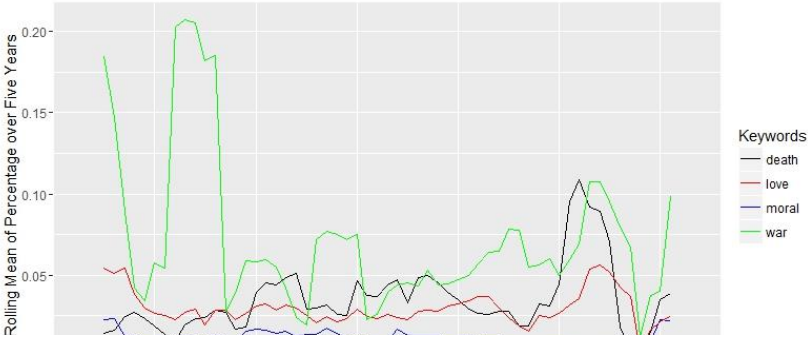

Figure 11. Rolling mean of frequencies of thematic keywords over five years from another collection of the search for "Dazai Osamu"

what made Japan initiate wars in Meiji Japan and World War II. Literature served as a lens for this inquiry. Even when the author's works did not directly involve wars, historians and literary critics read between the lines, trying to find hints of militant ideology. Figure 11 is a similar graph to Figure 10, but based on 470 documents from DfR that mention "Dazai Osamu". The graph shows a similar interest of wars in Dazai's works, although most of them did not tell war stories. Whether the author wrote about war was not important, western scholars were interested in finding the connections between literature and war if the author was famous in the late 19th and early 20th century Japan, when Sino-Japanese War, the Russo-Japanese War and World War II took place. Nevertheless, Soseki's connection to war was not unfound, since in his essay *Mankan Tokorodokoro (Travels in Manchuria and Korea)*, Soseki justified Japanese colonization in Manchuria and Korea.[6] Soseki thought that imperialism, colonization and wars were unavoidable consequences for a nation in its effort of modernization. Japan was only following steps of western colonizing powers.

The historiographical research on Natsume Soseki with DfR reveals the increased interests in Soseki and Japanese literature in general after 1950 in English academia. While the early studies focused on critics of Soseki's works and their translation, scholars became interested in situating Soseki in East Asian studies, especially in historical contexts. Although Soseki's novels had limited connections to wars, English scholars tended to emphasize the military implications, not only in Soseki's work, but also in Japanese literature in general. Japanese role as an Axis power in

6 Natsume, Brodey and Tsunematsu, *Rediscovering*, 1-30.

World War II and Soseki's ideological connections between modernity and warfare justified the emphasis. In next part of the paper, text mining of Soseki's fifteen novels also provides partial answer for the question of emphasis on war in English academia.

Text Mining of Fifteen Novels of Natsume Soseki

Soseki published fifteen novels between 1905 and 1916. Chronologically, they were *Wagahaiwa Nekodearu (I am a Cat)*, *Bocchan (Little Master)*, *Kusamakura (Pillow of Grass)*, *Nihyakutoka (The 210th day)*, *Nowaki (Autumn Wind)*, *Gubijinso (the Poppy)*, *Kofu (the Miner)*, *Sanshiro*, *Sorekara (And Then)*, *Mon (The Gate)*, *Higansugimade (Until after the Equinox), Kojin (The Wanderer)*, *Kokoro*, *Michikusa* and *Meian (Light and Darkness)*. This is not a corpus that too large to read; in fact, Soseki's critics read most of them. The character count is about 2.5 million, and the word count is about 1.5 million, with texted tokenized according to IPA dictionary in MeCab. Using quantitative method to study these fifteen texts, the paper points out some interesting patterns of Soseki's novels.

The texts of fifteen novels come from Aozora Bunko, a free Japanese digital library. The novels are in txt format with pronunciation guides and style annotations, which are cleaned out by R programs. The paper studies two statistics. The first is word frequency in percentage; the second

$$tf\text{-}idf = (1 + \ln(tf)) \times \ln\left(1 + \frac{|D|}{df + 1}\right)$$

Formula 1. Normalized tf-idf formula. tf stands for term frequency; df stands for document frequency; |D| stands for the size of documents.

is tf-idf. The formula for tf-idf used in this paper is Formula 1. Tf-idf is an index that measures the importance of a term in a document, but if the document size is not large, the inversed document frequency is not going to be small for common terms. In only fifteen Soseki's novels, words with high raw frequencies may also have a relatively high value of tf-idf index.

The historiographical research shows that English scholars were interested in connections between Soseki and wars. This interest is not unfound, since the character for military (軍) and for war (戦) appear in most of the novels, as shown in Figure12. The novels that do not contain the character military and war are *Nihyakutoka* and *Kofu*. There is the problem that some irrelevant

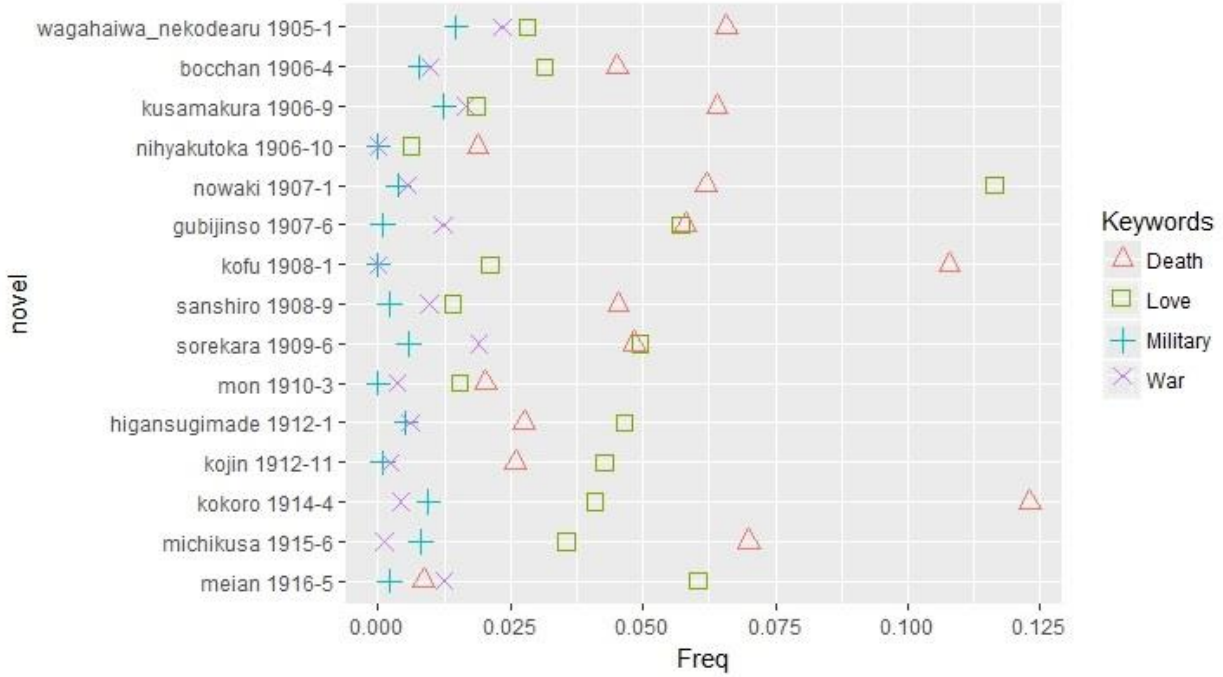

Figure 9. Frequencies in percentage of words that contain character 軍 (military), 戦 (war), 愛 (love) and 死 (death) in each novel

words contain the key character; for example, the search for the character love (愛) returns the word *Atago* (愛宕), which is a place name in Tokyo. These irrelevant words are excluded. The search for the character military (軍) yields results like army (陸軍), navy (海軍), naval ship (軍艦), general/shogun (将軍) and serviceman (軍人). They should arguably have different weight according to their connection to military, but in Figure 12, they are treated the same. Although the words related to war appear less than words contain love or death (死), Soseki's works were not devoid of war. In *Kokoro*, the father of Sensei's wife is a serviceman; In *Higansugimade*, Sunaga is a military family; In *Kusamakura*, there is a discussion of joining the army between siblings. Characters are closely related to wars, although Soseki and his family had little connections to the military world. Soseki's father was a *Nanushi* (village headman), and father-in-law was the chief secretary of the House of Peers.[7] Soseki had more connection to the Japanese nobles in administrative affairs, but in novels he also showed his interests in military organizations. The name of the war "Russo-Japanese" appears 13 times, and "Sino-Japanese", or more precisely "Japan-Qing" as in Japanese, appears 7 times in all the novels. Soseki's novels confirmed that military victories were important in shaping the modernizing Japan.

Nihyakutoka and *Kofu* are two exceptions from previous discussion, since both do not use words related to war or military. *Nihyakutoka* does not even frequently use some more common words like love and death because of its stylistic peculiarity. It is a novella, 15,000 words long, of conversations between two mountain climbers. *Kofu* tells the story about a rich young man,

7 McClellan, *Two*, 1-8

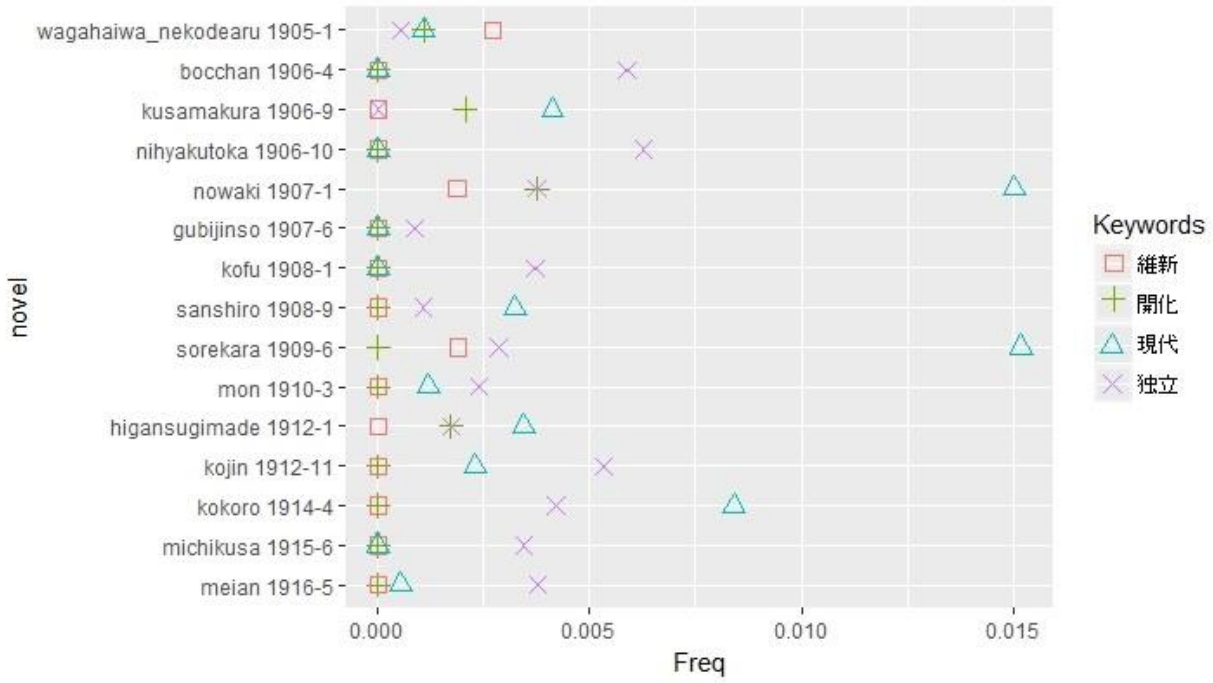

Figure 13. Frequencies in percentage of words related to modernity in each novel

exhausted by his uncomfortable relationship with his family and girlfriend, leaves his house in Tokyo, but is deceived into working as a miner. Not mentioning war in a work does not mean the absence of modernity. The encounter with modernity is the central theme of Soseki's works. In *Nihyakutoka*, the conversation between the two climbers leads to a discussion on Charles Dicekens, because prototypes of the two people are Soseki and his friend. In Kofu, the young man's experience is an epitome of the repressed working class in western word. Chozo, the man who deceives the young man, is a classical capitalist whose only interest is searching cheap labor. When the young man escapes the mine, he understands the dark side of capitalism.

An intuitive way to study how Soseki's novels dealt with modernity of tradition is to show the frequencies of directly related words. Figure 13 presents four words, restoration (維新), enlightenment (開化), modernity (現代) and independent (独立), directly related to modernity and

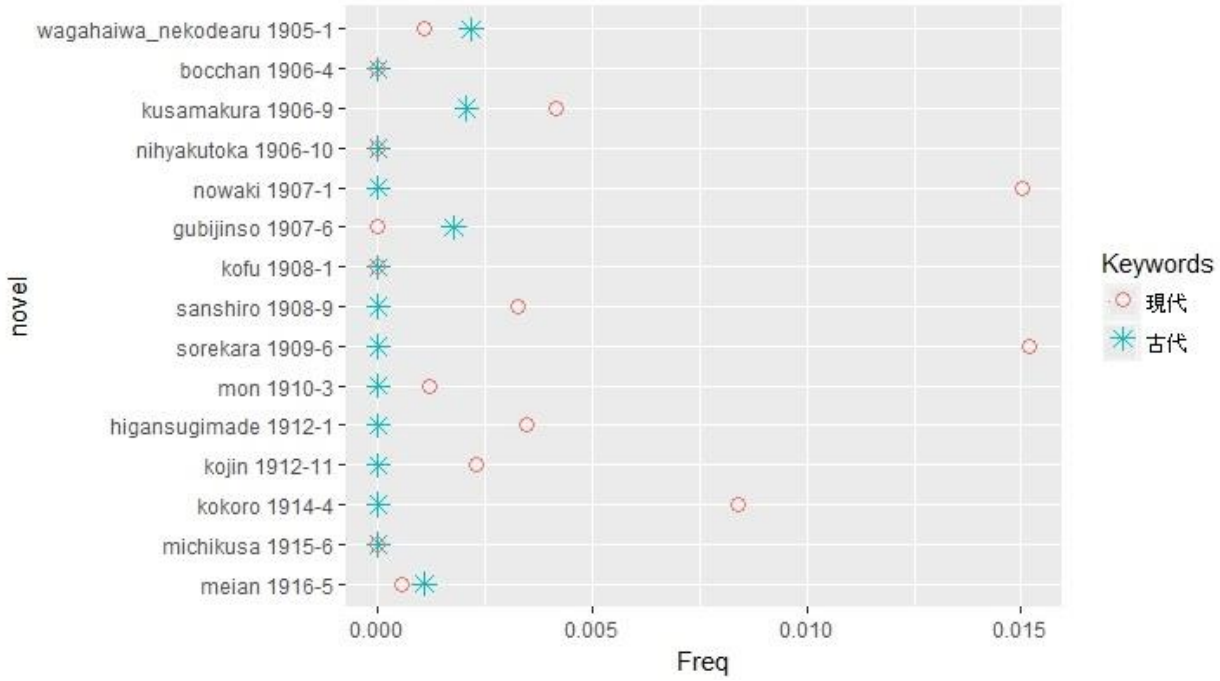

Figure 14. Frequencies in percentage of words modernity and antiquity in each novel

Meiji Restoration. Each work at least uses one of these words, and the average words a work uses is 2.01 out of 4. Figure 14 shows a comparison between words modernity and antiquity (古代); more works uses the word modernity. The figure does not contain words, such as tradition (伝統), conservatism (保守) and Japanese national learning (国学), since Soseki never used them. It might be the case that words like tradition or national learning emphasize superiority of Japanese culture, so Soseki avoided using them. His was disgusted by the shallow nationalist movement that extolled Japanese virtues when he was student.[8] The result in Figure 14 seems to suggest that Soseki was more interested in modernity, but a different perspective in Figure 15 shows a more ambivalent result. If Britain represented the modernity and China represented the tradition, Meiji Japan was the country in between. This graph illustrates how often did Soseki mentions Chinese and English

8 Ibid., 5-6.

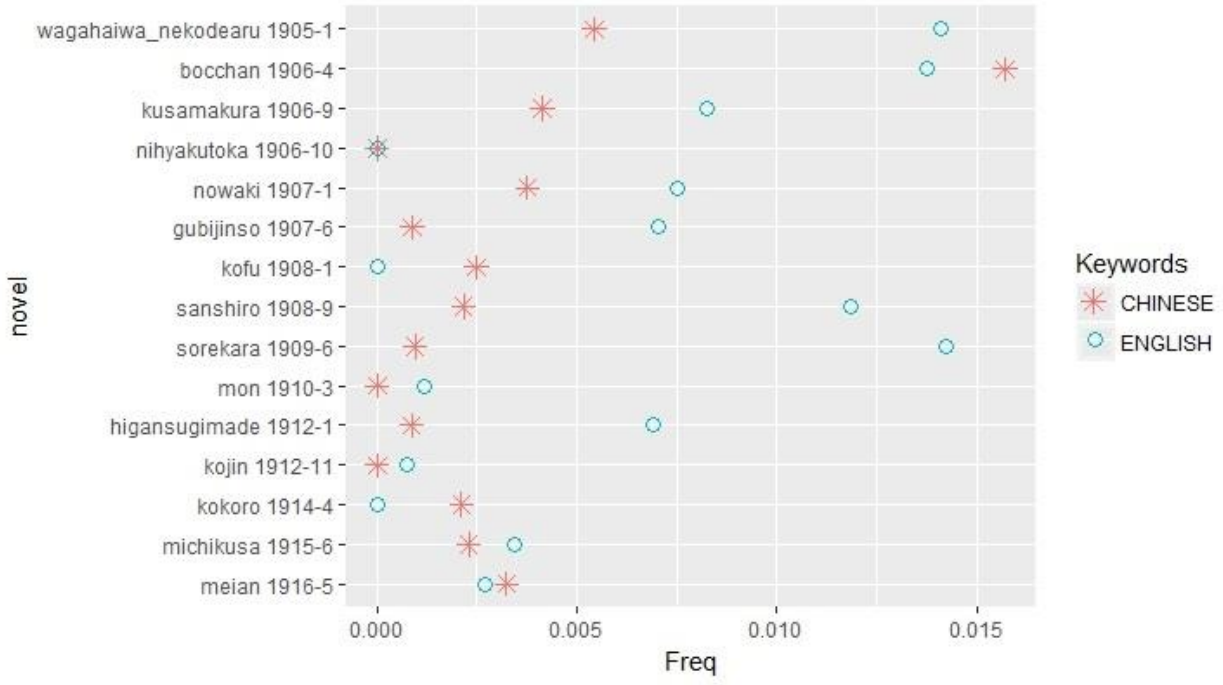

Figure 15. Frequencies in percentage of words related to Chinese and English in each novel

in his novels. To list a few, related words of Chinese include Qing Empire (清国), China (中国), Chinese book (漢籍) and Chinese poetry (漢詩); related words to English include the UK (イギリス/英国), English (英語/英文), Anglo-Japanese (英和) and English translation (英訳). Although the words related to English appear more than Chinese in ten out of fifteen of Soseki's works, the result is not as extreme as the comparison between modernity and antiquity. Since the westernizing Japan was in a fever of imitating the western countries, both culturally and politically, mentioning words related to English was not unusual for an English scholar like Soseki. Nevertheless, Soseki was more composed than fervent advocates of westernization and astutely aware of the pains in the modernization. During his two unpleasant years in London, Soseki showed his disdain for the ostensible gentleman manner of both Englishmen and Japanese in England. When his private tutor William J Craig, the editor the first *Oxford Shakespeare*,

commented that “no more than one in a hundred Englishmen would understand poetry”, Soseki shared this contempt.[9] Material success of the modernization did not necessarily result in intellectual superiority. Soseki was a supporter of westernization, because only powerful countries could resist invasion and colonization. Intellectually, however, Soseki rejected the superiority of western countries and was more inclined to the traditional Chinese culture. His *kanshi* rhymed in Chinese tones, which Japanese writers usually did not pay attention. In his last years, he rejected the professorship in English literature, finding consolation in Chinese paintings.[10] This inclination to the Chinese culture might come from mere personal preference or from the idea that origin of part of Japanese culture was Chinese culture, but it did not deny Japanese effort in modernization. Otherwise, the spirit of Meiji would not be a major theme of *Kokoro*.

A Close Look at *Kokoro*

Kokoro is one of Soseki’s most beloved novel. Figure 16 shows interesting pattern of the tf-idf of this novel. Most words in the graph for other novels are character names, since tf-idf index picks out distinctive words. *Kokoro*, however, do not have any terms with high tf-idf; the words from this novel crowd at the bottom of the graph. The word with the highest tf-idf, *Zoshigaya*, is a place name in Tokyo, while other words are related to the plot. McClellan commented that the reluctance to give characters names implies that Soseki wrote the novel as an “allegory of sorts”.

9 Ibid., 12.
10 Fukuchi, “*Kokoro*”, 480.

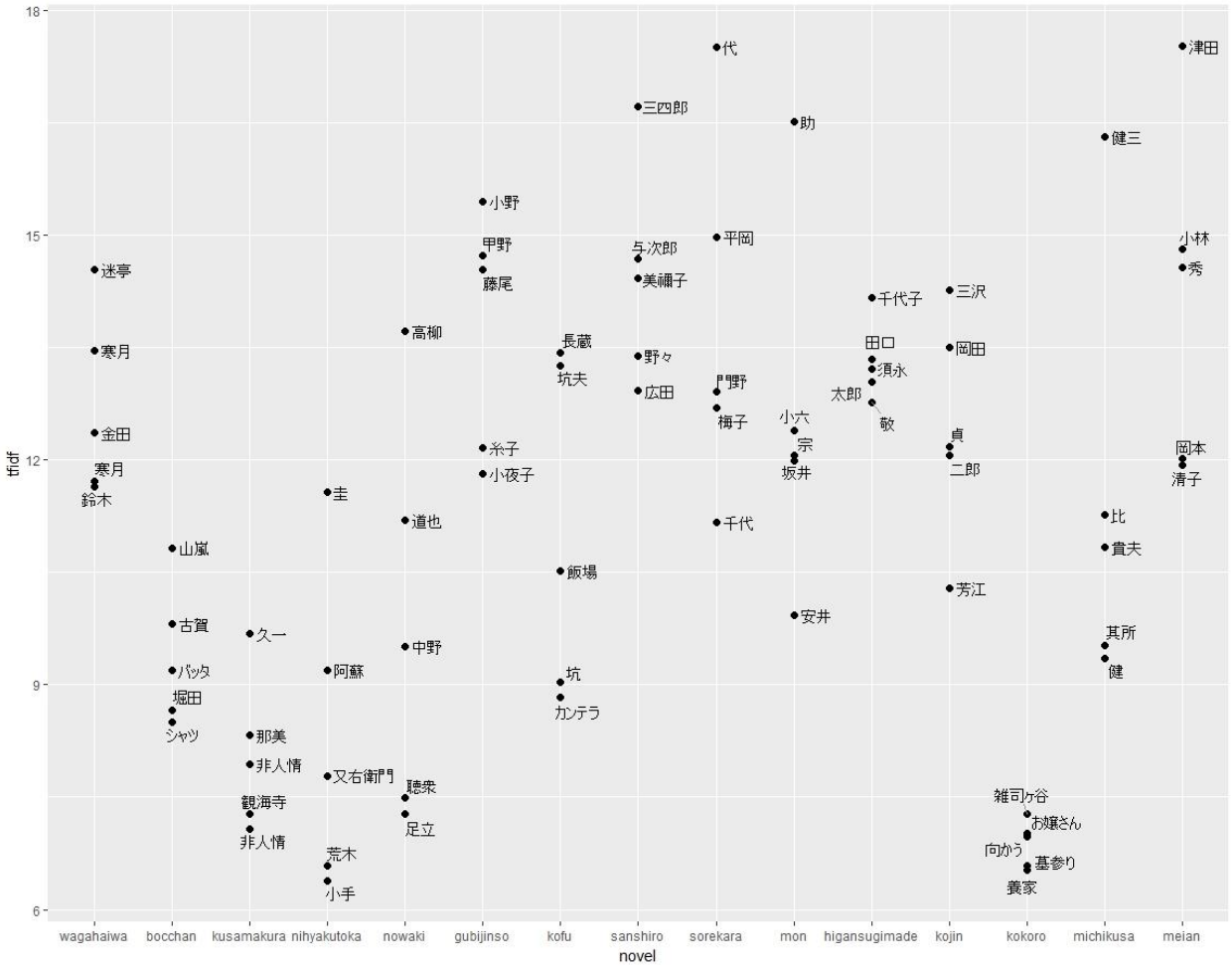

Figure 16. The top five tf-idf words in each novel

The loneliness and egoism of Sensei, one of the protagonists, could be common pain for any Meiji intellectuals. The novel was not only stylistically simple, but also lexically simple, as Figure 17 illustrates. While violins of the top 20 tf-idf words in other novels are stretched out, *Kokoro* has a short round shape because of its simple vocabulary. As previously mentioned, using tf-idf for a small corpse of documents does not eliminates high frequency words in novels. Figure 18 shows that in the top 20 words of *Kokoro*, nine out of twenty are common words with only one kana or character. For other Soseki's novels, top twenty tf-idf words contain less of these words.

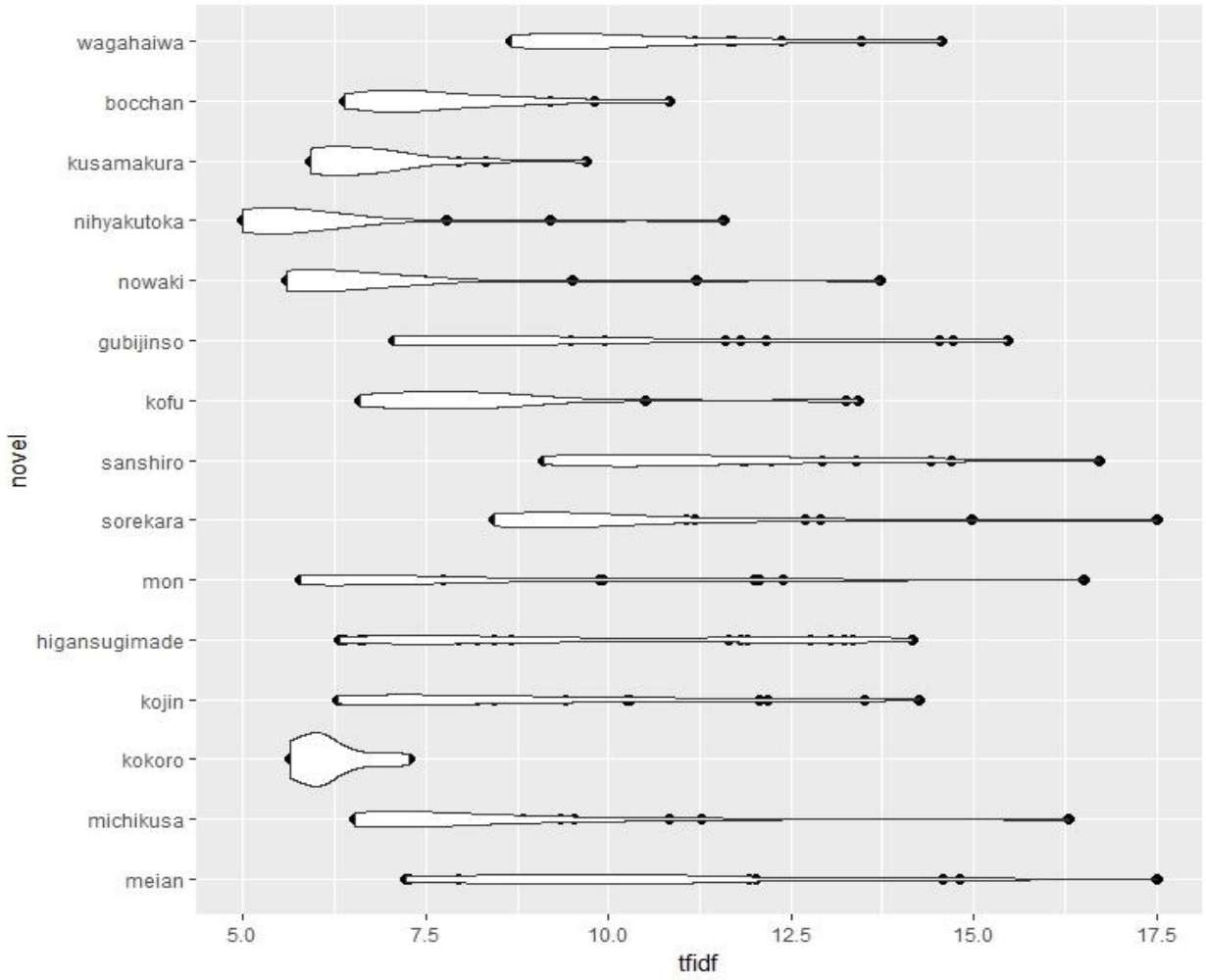

Figure 17. The top 20 tf-idf values for each novel

Kokoro appears to be a symbolic novel among Soseki's works. While other novels are not unimportant, *Kokoro* does present many themes that are common in all the novels. The book has three volumes. The young man is the protagonist in the first two volumes, while the third volume is a long letter from Sensei to the young man before Sensei commits suicide.

Literary critics paid much attention to the third volume, in which Soseki wrote about the relationship between the spirit of Meiji and Sensei's *junshi*, suicide through fidelity, to Meiji

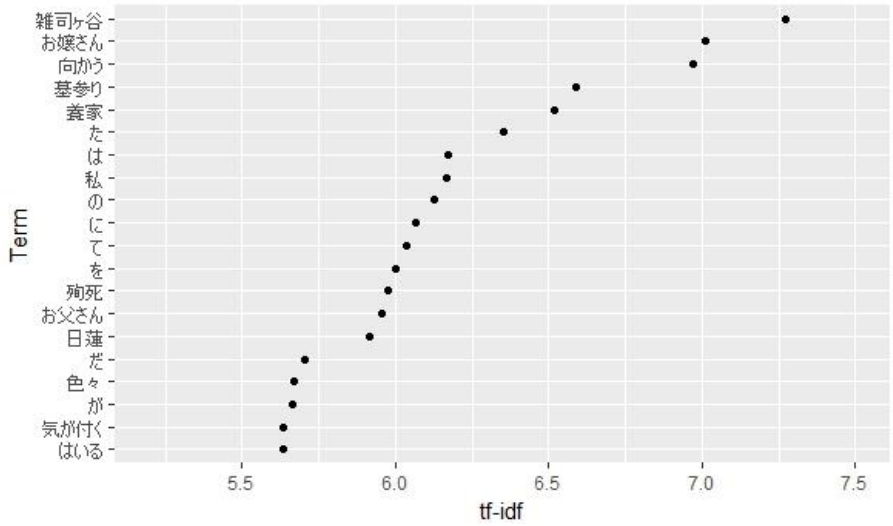

Figure 18. The top 20 tf-idf words for *Kokoro*

emperor. The volume is a response of Soseki to the end of the Meiji period. Sensei suffers from the loneliness and isolation that modernity brings to him, but he is a supporter of the modernity. His *junshi* to the emperor was an imitation of general Nogi Maresuke. Some of the Sensei's trauma comes from his awareness and regret of betrayal of his friend K, which leads to K's suicide. However, there is a transition from the feel of guilty to a more general sadness of being modern.[11] Sensei is a modern individual. He attends college, and some of his family members work in the army. Nevertheless, he is not happier than other people who are not modern. The word lonely (淋しい) composes about 0.03 percent of the total words of *Kokoro*, which is higher than any other Soseki's works. Loneliness summarizes the theme of *Kokoro*. Many literary critics quoted Sensei's comment that "loneliness is the price we pay to be modern".

The perspective of the first two volumes is that of the young man. The title of the first volume

11 Ibid., 483.

is “Sensei and I”, and the second is “parents and I”. Soseki built contradictions between Sensei and the young man’s father, who is a farmer. While Sensei represents a modern individual, the father represents a traditional one. Modernity does not bring happiness. When the young man reflects on the marriage between Sensei and his wife, he concludes that their relationship is a tragedy. Sensei hides his betrayal of K, but was inflicted in conscience. He could not tell this to his wife, since the betrayal is directed related to Sensei’s proposal. On the other hand, in the second chapter, the young man observes the sincere relationship between his parents. Whenever his dying father wants to see his mother, she stops her work and comes to the bed. They look at each other attentively and silently, until the father starts to cry and the mother tries to console him. The young man is impressed by the love of his parents, but is still determined to become a modern individual. Even when he looks after his ill father in his hometown, he keeps writing to Sensei in Tokyo. Although Sensei informs him of the trauma in modernity, the young man desires to leave his father and become an intellectual like Sensei. The young man’s determination to embrace modernity is close to that of Soseki. As a modern writer and scholar, Soseki was a supporter of the spirit of Meiji and modernization despite all of pain they brought.

Conclusion

The conclusions with text mining method and close reading are similar. Soseki was a proponent of modernity, although he was aware of the problems with rapid Japanese westernization. In the search of modernity and tradition in Soseki’s novels, the author was sometimes perplexed himself. Soseki maintained a close connection to traditional Chinese culture; his loved Chinese

literature and had talent in creating *kanshi* and paintings. Nonetheless, he was aware that modernization was the only option for a more powerful Japan.

Because Soseki was the youngest child, his parents sent him to be an adopted son in another family. The economic difficulty and the divorce of his adopted parents forced him to return to his original family, but Soseki's biological parents were never kind to him. His drifting between families and lack of love fostered Soseki's eccentric character. He found consolation in literature, through which he expressed love and loneliness in modern Japan. His two-year study in England nurtured his madness and solitude. The problems brought by modernity reflected in his novels came partly from the Soseki's personal experience. In *Kokoro*, Soseki illustrated that individualism would unexpectedly isolate a person. Sensei's guilt and isolation lead to his *junshi*. Soseki had similar experience in his own pursuit of modernity. He spent his happiest years with his wife in Kumamoto from 1896 to 1900. After his study in England, he became a mad man prone to domestic violence.[12] Soseki suffered from living in a modern country. Modernity was not a prerequisite of happiness. In his last years, he found consolation in novels and paintings. Although Soseki supported modernity, he was aware that aimless and simplistic imitation of western nations would result in pain.

While word frequency and measure of distinctive words help understand Soseki's support for modernization, quantitative historiographical research shows the interest of English scholars in the historical importance of Soseki's works. In studying this subject, the western academia paid

12 McClellan, *Two*, 9-10.

much attention to wars in literature. World War II and wars in Meiji Japan justified the focus. Text mining shows that although war was not a central theme in Soseki's works, it was relevant to many of his characters.

Distant reading cannot replace close reading, but it answers several questions quite well. What do scholars discuss when they study Soseki, what are some central themes in Soseki's works and how Soseki place himself between modernity and tradition? Some answers are surprising, such as western scholars' focus on war, while others agree with most literary critics, like Soseki's support for modernization. Soseki was so sensitive to the danger in rapid westernization that at some point he seemed to be confused in the search between modernity and tradition. Nevertheless, Soseki never denied the importance of modernization. His love for traditional culture did not stop him from admiring the Meiji spirit. Soseki was in nature a supporter of modernity.

Bibliography

Blevins, Cameron. “Space, Nation, and the Triumph of Region: A View of the World from Houston.” *Journal of American History* 101.1, (2014): 122-147.

Edelstein, Dan. “Enlightenment Scholarship by the Numbers: dfr.jstor.org, Dirty Quantification, and the Future of the Lit Review.” *Republic of Letters* Vol 4, Issue 1, (2014): 1-26.

Fukuchi, Isamu, “*Kokoro* and ‘the Spirit of Meiji,” *Monumenta Nipponica* 48.4, (1993): 469-488.

Natsume, Sōseki, Brodey, Inger and Tsunematsu, Sammy. *Rediscovering, Natsume Sōseki: with the first English translation of Travels in Manchuria and Korea; celebrating the centenary of Soseki's arrival in England, 1900 - 1902.* Folkestone: Global Oriental, 2000.

McClellan, Edwin. *Two Japanese novelists: Sōseki and Tōson.* Chicago: University of Chicago Press, 1969.

Appendix

R Code

A List of packages

```
library(stringr)
library(ggplot2)
library(data.table)
library(ggrepel)
library(RMeCab)
library(plyr)
library(tidyr)
library(zoo)
```

Windows users may want to use this to avoid problems in encoding

```
Sys.setlocale("LC_ALL", "Japanese")
```

This part of the code is for cleaning out annotations in the txt file downloaded from Aozora Bunko

```
file.org.dir<- list.files("D:/Google Drive/JPN_LIT/Natsume/original")
file.dir <- paste("D:/Google Drive/JPN_LIT/Natsume/original/",list.files("D:/Google
Drive/JPN_LIT/Natsume/original"), sep ="")

file.i <- 1
for (file.i in 1:length(file.dir)){
        Text.df <- read.delim(file.dir[file.i], header = FALSE, stringsAsFactors = FALSE, encoding = "SHIFT-
JIS")
        Text.text <- paste(Text.df[,1],collapse = "")
        Text.splited.raw <- unlist(str_split(Text.text, pattern = ""))
        Text.splited <- str_replace_all(Text.splited.raw, "｜", "") # Take out all "｜"

        ## Take out ruby and style notation
        ## Find out where to start and end
        start <- grep(pattern = "《|［", Text.splited)
        end <- grep(pattern = "》|］", Text.splited)
        from <- end + 1
        to <- start - 1
        real.from <- c(1, from)
        real.to <- c(to, length(Text.splited))
```

```
        CUT.df <- data.frame("from" = real.from, "to"= real.to,"text" = NA)

        ## Solve the situation when form > end
        row <- 1
        CUT.fine.df    <- data.frame("from" = 0, "to" = 0, "text" = NA)
        for(row in 1:length(CUT.df$from)){
                if(CUT.df$from[row] <= CUT.df$to[row]){
                        CUT.fine.df<- rbind(CUT.fine.df, CUT.df[row,])
                }
        }

        i <- 1
        for(i in 1:length(CUT.fine.df$from)){
                text <- Text.splited[CUT.fine.df$from[i]:CUT.fine.df$to[i]]
                CUT.fine.df$text[i] <- paste(text, collapse = "")
        }

        Text.cleaned.text <- paste(CUT.fine.df$text, collapse = "")

        write.table(Text.cleaned.text,file.org.dir[file.i],row.names = FALSE, col.names = FALSE)
}
```

This part of code is for text mining and plotting

Use Japanese Tokenizer to make frequency dataframes

```
na.zzz <- RMeCabFreq("D:/Google Drive/JPN_LIT/Natsume/zzz.txt")
na.zzz.reduced <- na.zzz[which(na.zzz$Info1 != "記号"),]

files.cle.dir<- list.files("D:/Google Drive/JPN_LIT/Natsume/cleaned")

for (n.i in 1:length(files.cle.dir)){
        assign(paste0("n.", files.cle.dir[n.i]), RMeCabFreq(paste0("D:/Google
Drive/JPN_LIT/Natsume/cleaned/",files.cle.dir[n.i]))[which(RMeCabFreq(paste0("D:/Google
Drive/JPN_LIT/Natsume/cleaned/",files.cle.dir[n.i]))$Info1 != "記号"),])
}

na.zzz.reduced$bocchan <- 0
na.zzz.reduced$gubijinso <- 0
na.zzz.reduced$higansugimade <- 0
na.zzz.reduced$kofu <- 0
na.zzz.reduced$kojin <- 0
```

```
na.zzz.reduced$kokoro <- 0
na.zzz.reduced$kusamakura <- 0
na.zzz.reduced$meian <- 0
na.zzz.reduced$michikusa <- 0
na.zzz.reduced$mon <- 0
na.zzz.reduced$nihyakutoka <- 0
na.zzz.reduced$nowaki <- 0
na.zzz.reduced$sanshiro <- 0
na.zzz.reduced$sorekara <- 0
na.zzz.reduced$wagahaiwa_nekodearu <- 0
```

This part of the code fill in the frequency of terms in each novel to a total frequency table of 15 novels. Reuse the code every time or write a loop based on the code.

```
z <- 1
y <- 1
while (z <= length(na.zzz.reduced$Term)){
        if(n.bocchan.txt$Term[y] == na.zzz.reduced$Term[z] & n.bocchan.txt$Info1[y] ==
na.zzz.reduced$Info1[z] & n.bocchan.txt$Info2[y] == na.zzz.reduced$Info2[z]){
                na.zzz.reduced$bocchan[z] <- n.bocchan.txt$Freq[y]
                y <-    y + 1
        }
        z <- z + 1
}
```

A Final tf-idf table

```
n.TMD <- na.zzz.reduced[,4:19]
n.TMD$Dfreq <- apply(n.TMD, 1, function(x) length(which(x != 0)))
n.TMD$Dfnorm <- log(15/n.TMD$Dfreq +1)

n.TFIDF.df <- data.frame(t(apply(n.TMD[,2:16], 1, function(x) log(x)+1)))
n.TFIDF.df <- n.TFIDF.df*n.TMD$Dfnorm
n.TFIDF.df[n.TFIDF.df == -Inf]    <- 0

Final.TFIDF.df <- cbind(na.zzz.reduced[,1:3],n.TFIDF.df)
```

Make another table of Final percentage

```
n.PERC.df<- data.frame(apply(n.TMD[,1:16], 2,function(x) x/sum(x)*100))
Final.PERC.df <- cbind(na.zzz.reduced[,1:3],n.PERC.df)
```

```
ord.kokoro.TFIDF.df <- Final.TFIDF.df[order(-Final.TFIDF.df$kokoro),]
ord.bocchan.TFIDF.df <- Final.TFIDF.df[order(-Final.TFIDF.df[,4]),]

ord.kokoro.TFIDF20.df <- ord.kokoro.TFIDF.df[1:20,]

ord.kokoro.TFIDF20.df$novel <- "kokoro"
ord.kokoro.TFIDF20.df$temp <- 2


###Make a dataframe of top 20 tfidf for each novel
all.TFIDF20 <- data.frame(Term = NA,tfidf = NA, novel = NA)
for (col.i in 4:18){
        all.TFIDF20 <- rbindlist(list(all.TFIDF20, cbind(Final.TFIDF.df[order(-
Final.TFIDF.df[,col.i]),][1:20,][,c(1,col.i)],colnames(Final.TFIDF.df)[col.i])))
}
all.TFIDF20 <- all.TFIDF20[-1,]
all.TFIDF20$novel    <- as.character(all.TFIDF20$novel)
all.TFIDF20[all.TFIDF20 == "wagahaiwa_nekodearu"] <- "wagahaiwa"

all.TFIDF5 <- data.frame(Term = NA,tfidf = NA, novel = NA)
for (col.i in 4:18){
        all.TFIDF5 <- rbindlist(list(all.TFIDF5, cbind(Final.TFIDF.df[order(-
Final.TFIDF.df[,col.i]),][1:5,][,c(1,col.i)],as.character(colnames(Final.TFIDF.df)[col.i]))))
}
all.TFIDF5 <- all.TFIDF5[-1,]
all.TFIDF5$novel    <- as.character(all.TFIDF5$novel)
```

Plotting Part

```
ggplot(all.TFIDF20, aes(x = novel, y = tfidf))+
        geom_point(size = 2) +

scale_x_discrete(limits=c("wagahaiwa","bocchan","kusamakura","nihyakutoka","nowaki","gubijinso","kofu","sansh
iro","sorekara","mon","higansugimade","kojin","kokoro","michikusa","meian")) +
```

Violin Graph

```
ggplot(all.TFIDF20, aes(x = novel, y = tfidf))+
        geom_point(size = 2) +

scale_x_discrete(limits=rev(c("wagahaiwa","bocchan","kusamakura","nihyakutoka","nowaki","gubijinso","kofu","s
```

```
anshiro","sorekara","mon","higansugimade","kojin","kokoro","michikusa","meian"))) +
        geom_violin()+
        coord_flip()


## Top five tf-idf
all.TFIDF5[all.TFIDF5 == "wagahaiwa_nekodearu"] <- "wagahaiwa"
ggplot(all.TFIDF5, aes(x = novel, y = tfidf))+
        geom_point(size = 2) +

scale_x_discrete(limits=c("wagahaiwa","bocchan","kusamakura","nihyakutoka","nowaki","gubijinso","kofu","sansh
iro","sorekara","mon","higansugimade","kojin","kokoro","michikusa","meian")) +
        geom_text_repel(aes(label=Term), size =4, segment.color = 'grey60',nudge_x =0.05)

ggplot(subset(Final.TFIDF.df, Term != "ない" & Info1 == ",形容詞" & kokoro >=3),aes(x = kokoro, y =
reorder(Term, kokoro))) +
        geom_point()

### Frequency of military, war love and death
### Gun
gun.Freq.df <- Final.PERC.df[grep(pattern = "軍", Final.TFIDF.df$Term),]
gun.Freq.df <- data.frame(apply(gun.Freq.df[,5:19], 2, function(x) sum(x)))
gun.Freq.df$novel <- row.names(gun.Freq.df)
colnames(gun.Freq.df)[1] <- "Freq"

### Sen
sen.Freq.df <- Final.PERC.df[grep(pattern = "戦", Final.TFIDF.df$Term),]
sen.Freq.df <- data.frame(apply(sen.Freq.df[,5:19], 2, function(x) sum(x)))
sen.Freq.df$novel <- row.names(sen.Freq.df)
colnames(sen.Freq.df)[1] <- "Freq"

### Ai
ai.Freq.df <- Final.PERC.df[grep(pattern = "愛", Final.TFIDF.df$Term),]
ai.Freq.df <- ai.Freq.df[which(ai.Freq.df$Term != "愛宕" ),]
ai.Freq.df <- ai.Freq.df[which(ai.Freq.df$Term != "愛宕山" ),]
ai.Freq.df <- data.frame(apply(ai.Freq.df[,5:19], 2, function(x) sum(x)))
ai.Freq.df$novel <- row.names(ai.Freq.df)
colnames(ai.Freq.df)[1] <- "Freq"

### Shi
shi.Freq.df <- Final.PERC.df[grep(pattern = "死", Final.TFIDF.df$Term),]
```

```
shi.Freq.df <- data.frame(apply(shi.Freq.df[,5:19], 2, function(x) sum(x)))
shi.Freq.df$novel <- row.names(shi.Freq.df)
colnames(shi.Freq.df)[1] <- "Freq"

gunsenall.Freq.df <- cbind(gun.Freq.df, sen.Freq.df$Freq, ai.Freq.df$Freq, shi.Freq.df$Freq)
colnames(gunsenall.Freq.df)[3:5] <- c("sen","ai","shi")
gunsenall.Freq.df$novel <- c("bocchan 1906-4",
                             "gubijinso 1907-6",
                             "higansugimade 1912-1",
                             "kofu 1908-1",
                             "kojin 1912-11",
                             "kokoro 1914-4",
                             "kusamakura 1906-9",
                             "meian 1916-5",
                             "michikusa 1915-6",
                             "mon 1910-3",
                             "nihyakutoka 1906-10",
                             "nowaki 1907-1",
                             "sanshiro 1908-9",
                             "sorekara 1909-6",
                             "wagahaiwa_nekodearu 1905-1")
ggplot(gunsenall.Freq.df, aes(x=novel))+
        geom_point(aes(y = Freq, color = "Military"), shape =3, size = 3)+
        geom_point(aes(y = sen, color = "War"), shape =4, size = 3)+
        geom_point(aes(y = ai, color = "Love"), shape =0, size = 3)+
        geom_point(aes(y = shi, color = "Death"), shape =2, size = 3)+
        scale_x_discrete(limits=rev(c("wagahaiwa_nekodearu 1905-1",
                                      "bocchan 1906-4",
                                      "kusamakura 1906-9",
                                      "nihyakutoka 1906-10",
                                      "nowaki 1907-1",
                                      "gubijinso 1907-6",
                                      "kofu 1908-1",
                                      "sanshiro 1908-9",
                                      "sorekara 1909-6",
                                      "mon 1910-3",
                                      "higansugimade 1912-1",
                                      "kojin 1912-11",
                                      "kokoro 1914-4",
                                      "michikusa 1915-6",
                                      "meian 1916-5"))) +
```

```
        labs(color="Keywords")+
        coord_flip()
```

Graph of frequency of modernity and antiquity

```
### Gendai & Kodai
gendai.all.PERC.df <- Final.PERC.df[which(Final.PERC.df$Term == "維新"| Final.PERC.df$Term == "現代"
|Final.PERC.df$Term =="開化"|Final.PERC.df$Term =="独立"),]
gendai.all.PERC.df <- gendai.all.PERC.df[,-1:-4]
gendai.all.PERC.df[2,8] <- gendai.all.PERC.df[2,8] + gendai.all.PERC.df[5,8]
gendai.all.PERC.df <- gendai.all.PERC.df[1:4,]
t.gendai.df <- data.frame(t(gendai.all.PERC.df))
colnames(t.gendai.df) <- c("開化", "独立", "維新", "現代")
t.gendai.df$novel <- row.names(t.gendai.df)

t.gendai.df$novel <- c("bocchan 1906-4",
                       "gubijinso 1907-6",
                       "higansugimade 1912-1",
                       "kofu 1908-1",
                       "kojin 1912-11",
                       "kokoro 1914-4",
                       "kusamakura 1906-9",
                       "meian 1916-5",
                       "michikusa 1915-6",
                       "mon 1910-3",
                       "nihyakutoka 1906-10",
                       "nowaki 1907-1",
                       "sanshiro 1908-9",
                       "sorekara 1909-6",
                       "wagahaiwa_nekodearu 1905-1")


ggplot(t.gendai.df, aes(x = novel))+
        geom_point(aes(y=開化, color ="開化"), shape = 3, size =3) +
        geom_point(aes(y=独立, color = "独立"), shape = 4, size =3) +
        geom_point(aes(y=維新, color = "維新"), shape = 0, size =3) +
        geom_point(aes(y=現代, color = "現代"), shape = 2, size =3) +
        scale_x_discrete(limits=rev(c("wagahaiwa_nekodearu 1905-1",
                                      "bocchan 1906-4",
                                      "kusamakura 1906-9",
```

```
                                        "nihyakutoka 1906-10",
                                        "nowaki 1907-1",
                                        "gubijinso 1907-6",
                                        "kofu 1908-1",
                                        "sanshiro 1908-9",
                                        "sorekara 1909-6",
                                        "mon 1910-3",
                                        "higansugimade 1912-1",
                                        "kojin 1912-11",
                                        "kokoro 1914-4",
                                        "michikusa 1915-6",
                                        "meian 1916-5"))) +
        labs(color="Keywords")+
        ylab("Freq") +
        coord_flip()

gvk.all.PERC.df <- Final.PERC.df[which(Final.PERC.df$Term == "現代" |Final.PERC.df$Term =="古代"),]
gvk.all.PERC.df <- gvk.all.PERC.df[,-1:-4]
t.gvk.df <- data.frame(t(gvk.all.PERC.df))
colnames(t.gvk.df) <- c("現代", "古代")
t.gvk.df$novel <- row.names(t.gvk.df)

t.gvk.df$novel <- c("bocchan 1906-4",
                    "gubijinso 1907-6",
                    "higansugimade 1912-1",
                    "kofu 1908-1",
                    "kojin 1912-11",
                    "kokoro 1914-4",
                    "kusamakura 1906-9",
                    "meian 1916-5",
                    "michikusa 1915-6",
                    "mon 1910-3",
                    "nihyakutoka 1906-10",
                    "nowaki 1907-1",
                    "sanshiro 1908-9",
                    "sorekara 1909-6",
                    "wagahaiwa_nekodearu 1905-1")

ggplot(t.gvk.df, aes(x = novel))+
        geom_point(aes(y=古代, color = "古代"), shape = 8, size =3) +
        geom_point(aes(y=現代, color = "現代"), shape = 1, size =3) +
```

```
        scale_x_discrete(limits=rev(c("wagahaiwa_nekodearu 1905-1",
                                      "bocchan 1906-4",
                                      "kusamakura 1906-9",
                                      "nihyakutoka 1906-10",
                                      "nowaki 1907-1",
                                      "gubijinso 1907-6",
                                      "kofu 1908-1",
                                      "sanshiro 1908-9",
                                      "sorekara 1909-6",
                                      "mon 1910-3",
                                      "higansugimade 1912-1",
                                      "kojin 1912-11",
                                      "kokoro 1914-4",
                                      "michikusa 1915-6",
                                      "meian 1916-5"))) +
        labs(color="Keywords")+
        ylab("Freq") +
        coord_flip()
```

Graph of frequency of Chinese and English

```
chn.all.PERC.df <- Final.PERC.df[which(Final.PERC.df$Term == "清国" |
                                       Final.PERC.df$Term == "中国"|
                                       Final.PERC.df$Term == "漢学"|
                                       Final.PERC.df$Term =="漢語"|
                                       Final.PERC.df$Term =="漢詩"|
                                       Final.PERC.df$Term =="漢籍"|
                                       Final.PERC.df$Term =="漢土"|
                                       Final.PERC.df$Term =="漢"|
                                       Final.PERC.df$Term =="漢人"),]

eng.all.PERC.df <- Final.PERC.df[which(Final.PERC.df$Term == "イギリス" |
                                         Final.PERC.df$Term == "英国"|
                                         Final.PERC.df$Term == "英訳"|
                                         Final.PERC.df$Term =="英語"|
                                         Final.PERC.df$Term =="英文"|
                                         Final.PERC.df$Term =="英和"),]
chn.eng.df <- data.frame(apply(chn.all.PERC.df[,5:19], 2, function(x) sum(x)))
chn.eng.df <- cbind(chn.eng.df, data.frame(apply(eng.all.PERC.df[,5:19], 2, function(x) sum(x))))
```

```
colnames(chn.eng.df) <- c("CHINESE", "ENGLISH")

chn.eng.df$novel <- row.names(chn.eng.df)

chn.eng.df$novel <- c("bocchan 1906-4",
                      "gubijinso 1907-6",
                      "higansugimade 1912-1",
                      "kofu 1908-1",
                      "kojin 1912-11",
                      "kokoro 1914-4",
                      "kusamakura 1906-9",
                      "meian 1916-5",
                      "michikusa 1915-6",
                      "mon 1910-3",
                      "nihyakutoka 1906-10",
                      "nowaki 1907-1",
                      "sanshiro 1908-9",
                      "sorekara 1909-6",
                      "wagahaiwa_nekodearu 1905-1")

ggplot(chn.eng.df, aes(x = novel))+
        geom_point(aes(y=CHINESE, color = "CHINESE"), shape = 8, size =3) +
        geom_point(aes(y=ENGLISH, color = "ENGLISH"), shape = 1, size =3) +
        scale_x_discrete(limits=rev(c("wagahaiwa_nekodearu 1905-1",
                                      "bocchan 1906-4",
                                      "kusamakura 1906-9",
                                      "nihyakutoka 1906-10",
                                      "nowaki 1907-1",
                                      "gubijinso 1907-6",
                                      "kofu 1908-1",
                                      "sanshiro 1908-9",
                                      "sorekara 1909-6",
                                      "mon 1910-3",
                                      "higansugimade 1912-1",
                                      "kojin 1912-11",
                                      "kokoro 1914-4",
                                      "michikusa 1915-6",
                                      "meian 1916-5"))) +
        labs(color="Keywords")+
        ylab("Freq") +
        coord_flip()
```

```
ggplot(ord.kokoro.TFIDF20.df, aes(x = kokoro, y = reorder(Term, kokoro) ))+
        geom_point() +
        xlim(5.2,7.5) +
        ylab("Term") +
        xlab("tf-idf")
```

This part of the code is for historiographical research. The code is the same as the code given in class except for the plotting part.

```
##read in the file with the metadata
citations <- read.delim("C:/Users/klijia/Desktop/HIST582A/JSTOR/Dazai Osamu/citations1.txt",
stringsAsFactors=FALSE,encoding = "UTF-8")

##extract the year
citations$pubyear <- str_extract(citations$pubdate, "^¥¥d{4}")

##set folder name as variable
x_name <- "C:/Users/klijia/Desktop/HIST582A/JSTOR/Dazai Osamu/wordcounts"

##get all files in that folder
files <- list.files(path = x_name)
setwd("C:/Users/klijia/Desktop/HIST582A/JSTOR/Dazai Osamu/wordcounts")


##creates a doi variable from the file name by removing prefix and suffix
word.counter <- function(x){
WORD_counter.df <- read.table(x, sep=",", stringsAsFactors=FALSE, header=TRUE)
WORD_counter.df$doi <- str_replace(x, "wordcounts_", "")
WORD_counter.df$doi <- str_replace(WORD_counter.df$doi, ".CSV", "")
WORD_counter.df$doi <- str_replace(WORD_counter.df$doi, "_", "/")
return <- WORD_counter.df
}



Tokugawa.list <- lapply(files, word.counter)
Tokugawa.list[1]

Tokugawa.df <- rbindlist(Tokugawa.list)
```

```
##map the citation information in
colnames(Tokugawa.df) <- c("word","count","doi")

Tokugawa.df$pubyear <- Tokugawa.df$doi
Tokugawa.df$pubyear <- mapvalues(Tokugawa.df$pubyear, from=citations$doi, to=citations$pubyear)
Tokugawa.df$pubyear <- as.numeric(Tokugawa.df$pubyear)

Tokugawa.df$title <- Tokugawa.df$doi
Tokugawa.df$title <- mapvalues(Tokugawa.df$title, from=citations$doi, to=citations$title)




Tokugawa.agg.df <- aggregate(Tokugawa.df$count, by=list(Tokugawa.df$pubyear, Tokugawa.df$word), FUN=sum,
na.rm=TRUE)
colnames(Tokugawa.agg.df) <- c("pubyear","word","count")
Tokugawa.agg.df$pubyear <- as.numeric(Tokugawa.agg.df$pubyear)



##fill = 0 gives 0 instead of NA
Tokugawa.agg.df <- spread(Tokugawa.agg.df, key=word, value=count, fill=0)

ggplot(data = Tokugawa.agg.df, aes(pubyear, translation)) + geom_point()
ggplot(data = Tokugawa.agg.df, aes(pubyear, shakespeare)) + geom_point()
ggplot(data = Tokugawa.agg.df, aes(pubyear, seidensticker)) + geom_point()
ggplot(data = Tokugawa.agg.df, aes(pubyear, keene)) + geom_point()
ggplot(data = Tokugawa.agg.df, aes(pubyear, mcclellan)) + geom_point()
ggplot(data = Tokugawa.agg.df, aes(pubyear, mclinney)) + geom_point()
ggplot(data = Tokugawa.agg.df, aes(pubyear, a)) + geom_point()
ggplot() + geom_line(data = Tokugawa.agg.df, aes(pubyear, dazai), color = "red") +
          geom_line(data = Tokugawa.agg.df, aes(pubyear, kawabata), color = "blue") +
          geom_line(data = Tokugawa.agg.df, aes(pubyear, natsume), color = "green") +
          geom_line(data = Tokugawa.agg.df, aes(pubyear, mishima), color = "orange") +
          geom_line(data = Tokugawa.agg.df, aes(pubyear, murasaki), color = "purple") +
          geom_line(data = Tokugawa.agg.df, aes(pubyear, matsuo), color = "black")
ggplot() +geom_point(data = Tokugawa.agg.df, aes(pubyear, natsume), color = "green")

rowSums(Tokugawa.agg.df[c(2:ncol(Tokugawa.agg.df))])
```

```
## Get a word frequency table from the agg table
new.df    <- Tokugawa.agg.df
new.df    <- subset(new.df, select = -pubyear)
new.df <- data.frame(t(new.df))
new.df$sum <- rowSums(new.df)

Tokugawa.agg.perc.df <- Tokugawa.agg.df[,-1]

Tokugawa.agg.perc.df <- apply(X = Tokugawa.agg.perc.df, MARGIN = 1, FUN = function (x) x/sum(x))
Tokugawa.agg.perc.df <- t(Tokugawa.agg.perc.df*100)
Tokugawa.agg.perc.df <- data.frame(Tokugawa.agg.perc.df)

Tokugawa.agg.perc.df <- cbind.data.frame(Tokugawa.agg.df[,1], Tokugawa.agg.perc.df)
colnames(Tokugawa.agg.perc.df)[1] <- "pubyear"

# Tokugawa.agg.perc.df <-Tokugawa.agg.df[,-1]
##get rolling averages, but what about missing years?
full_seq(Tokugawa.agg.perc.df$pubyear,1) ##trick for sequence

Complete.year.df <- data.frame("pubyear"=full_seq(Tokugawa.agg.perc.df$pubyear,1))
Tokugawa.full.perc.df <- merge(Tokugawa.agg.perc.df, Complete.year.df, by = "pubyear", all = TRUE)
```

###The following codes are codes for each line graph in the historiographical research

```
keepers <- c("translation","seidensticker","keene","mcclellan")
Tokugawa.full.smaller <- Tokugawa.full.perc.df[,keepers]



Tokugawa.full.smaller[is.na(Tokugawa.full.smaller)] <- 0

Tokugawa.smaller.roll.5 <- data.frame(rollmean(Tokugawa.full.smaller, k=5, fill = list(NA, NULL, NA)))



Tokugawa.smaller.roll.5$pubyear <- Tokugawa.full.perc.df$pubyear

mathching <- c("translation" = "black","seidensticker" = "red","keene" = "blue","mcclellan" = "green")
ggplot(Tokugawa.smaller.roll.5, aes(x=pubyear)) +
        geom_line(aes(y = translation, color = "translation")) +
```

```
        geom_line(aes(y = seidensticker, color = "seidensticker"))+
        geom_line(aes(y = keene, color = "keene"))+
        geom_line(aes(y = mcclellan, color = "mcclellan")) +
        scale_colour_manual(name="Keywords",values = mathching)+
        xlab("Year") + ylab("Rolling Mean of Percentage over Five Years") +
        guides(col = guide_legend(reverse = TRUE))

##################################################


keepers <- c("japanese","american","english","chinese","korean")
Tokugawa.full.smaller <- Tokugawa.full.perc.df[,keepers]




ptm <- proc.time()
Tokugawa.full.smaller[is.na(Tokugawa.full.smaller)] <- 0
run_time <- proc.time() - ptm
run_time

Tokugawa.smaller.roll.5 <- data.frame(rollmean(Tokugawa.full.smaller, k=5, fill = list(NA, NULL, NA)))


Tokugawa.smaller.roll.5$pubyear <- Tokugawa.full.perc.df$pubyear
mathching <- c("japanese" = "black","english" = "blue","chinese" = "red","korean" = "green")
ggplot(Tokugawa.smaller.roll.5, aes(x=pubyear)) +
        geom_line(aes(y = japanese, color = "japanese")) +
        geom_line(aes(y = english, color = "english"))+
        geom_line(aes(y = chinese, color = "chinese")) +
        geom_line(aes(y = korean, color = "korean")) +
        scale_colour_manual(name="Keywords",values = mathching)+
        xlab("Year") + ylab("Rolling Mean of Percentage over Five Years") +
        guides(col = guide_legend(reverse = FALSE))

##################################################

keepers <- c("war","moral","love","death")
Tokugawa.full.smaller <- Tokugawa.full.perc.df[,keepers]

Tokugawa.full.smaller[is.na(Tokugawa.full.smaller)] <- 0
```

```
Tokugawa.smaller.roll.5 <- data.frame(rollmean(Tokugawa.full.smaller, k=5, fill = list(NA, NULL, NA)))


Tokugawa.smaller.roll.5$pubyear <- Tokugawa.full.perc.df$pubyear

mathching <- c("death" = "black","moral" = "blue","love" = "red","war" = "green")
ggplot(Tokugawa.smaller.roll.5, aes(x=pubyear)) +
        geom_line(aes(y = death, color = "death")) +
        geom_line(aes(y = moral, color = "moral"))+
        geom_line(aes(y = love, color = "love")) +
        geom_line(aes(y = war, color = "war")) +
        scale_colour_manual(name="Keywords",values = mathching)+
        xlab("Year") + ylab("Rolling Mean of Percentage over Five Years") +
        guides(col = guide_legend(reverse = FALSE))

############################################################


keepers <- c("social","political","economic","historical")
Tokugawa.full.smaller <- Tokugawa.full.perc.df[,keepers]

Tokugawa.full.smaller[is.na(Tokugawa.full.smaller)] <- 0


Tokugawa.smaller.roll.5 <- data.frame(rollmean(Tokugawa.full.smaller, k=5, fill = list(NA, NULL, NA)))


Tokugawa.smaller.roll.5$pubyear <- Tokugawa.full.perc.df$pubyear

mathching <- c("social" = "black","political" = "blue","economic" = "red","historical" = "green")
ggplot(Tokugawa.smaller.roll.5, aes(x=pubyear)) +
        geom_line(aes(y = social, color = "social")) +
        geom_line(aes(y = political, color = "political"))+
        geom_line(aes(y = economic, color = "economic")) +
        geom_line(aes(y = historical, color = "historical")) +
        scale_colour_manual(name="Keywords",values = mathching)+
        xlab("Year") + ylab("Rolling Mean of Percentage over Five Years") +
        guides(col = guide_legend(reverse = FALSE))
```

```
############################################################


keepers <- c("natsume","dazai","tanizaki","akutagawa","kawabata","matsuo","chikamatsu","murasaki")
Tokugawa.full.smaller <- Tokugawa.full.perc.df[,keepers]

Tokugawa.full.smaller[is.na(Tokugawa.full.smaller)] <- 0


Tokugawa.smaller.roll.5 <- data.frame(rollmean(Tokugawa.full.smaller, k=5, fill = list(NA, NULL, NA)))


Tokugawa.smaller.roll.5$pubyear <- Tokugawa.full.perc.df$pubyear

mathching <- c("natsume" = "red4","dazai" = "black","tanizaki" = "orange4","akutagawa" = "orange", "kawabata" =
"red", "matsuo" = "seagreen4","chikamatsu" = "palegreen","murasaki" = "royalblue")
ggplot(Tokugawa.smaller.roll.5, aes(x=pubyear)) +
        geom_line(aes(y = natsume, color = "natsume")) +
        geom_line(aes(y = dazai, color = "dazai"))+
        geom_line(aes(y = tanizaki, color = "tanizaki")) +
        geom_line(aes(y = akutagawa, color = "akutagawa")) +
        geom_line(aes(y = kawabata, color = "kawabata")) +
        geom_line(aes(y = matsuo, color = "matsuo")) +
        geom_line(aes(y = chikamatsu, color = "chikamatsu")) +
        geom_line(aes(y = murasaki, color = "murasaki")) +
        scale_colour_manual(name="Keywords",values = mathching)+
        xlab("Year") + ylab("Rolling Mean of Percentage over Five Years") +
        guides(col = guide_legend(reverse = FALSE))
```